Statistical Prediction Models for Network Traffic Performance

Kejia Hu, Alex Sim

Scientific Data Management Research Group
Computational Research Division
Lawrence Berkeley National Laboratory

AND

Demetris Antoniades, Constantine Dovrolis

College of Computing
Georgia Institute of Technology

Outline

- Goal
- Data
- Models
- Results

Goal

Statistical Prediction Model

- By analyzing
  - network traffic patterns and variation with the network conditions
  - based on two types of historical network measurement data
- To develop
  - performance prediction models for high-bandwidth networks
  - forecast the future network usage for a given time window and with a given probabilistic error requirement

Data and Modeling

- **SNMP Data**
  - **Single time series**
  - **With even collecting frequency**
  - **Develop time series model with Seasonal Adjustment**

System Time	Bytes
1336350930	7899633.733
1336350960	10665164.133
1336350990	13223715.5
1336351020	12133668.13
1336351050	12647888.6

Data and Modeling

- **Netflow Data**
  - **Multivariate data**
  - **Random colleting behavior**
  - **Multiple time series**
  - **Develop Generalized Linear Mixed Model**

Start time	0930.23:59:37.925
End time	0930.23:59:37.925
Source interface	179
Source IP address	xx.xx.xx.xx
Source Port	xxxxx
Destination Interface	175
Destination IP address	xxx.xxx.xxx.xxx
Destination Port	xxxxx
Packets	1
Octets	52

Modeling - Seasonal Adjustment

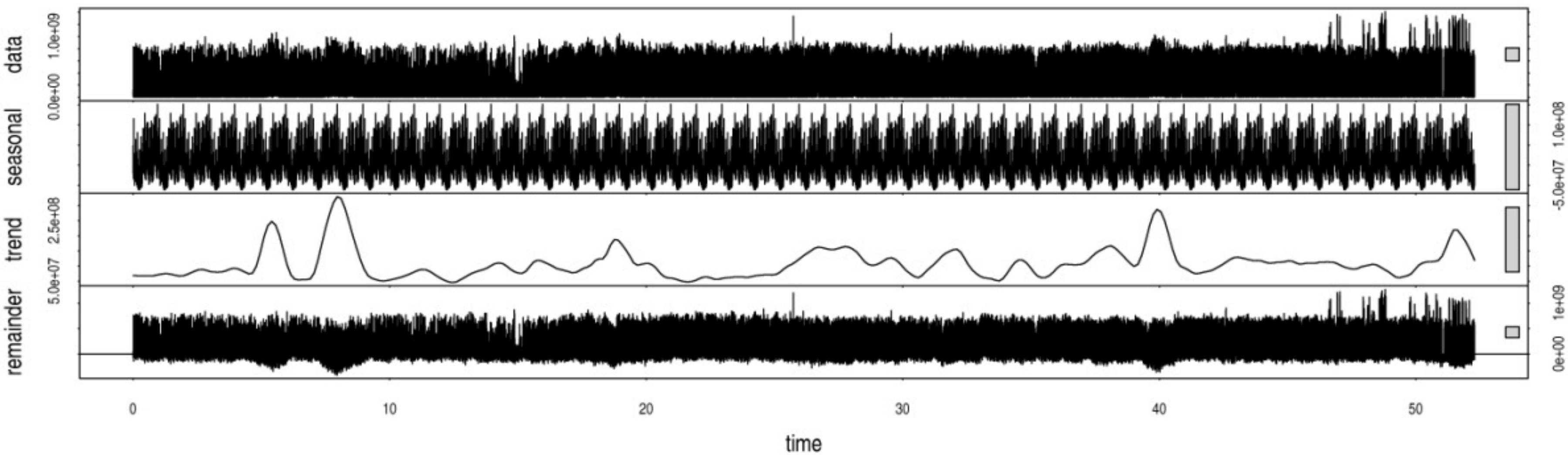
- A time series model with Seasonal Adjustment
- Decompose the SNMP data into three components
  - Seasonal component
  - Trend component
  - Random component.
- Enables an effective analysis in
  - prediction
  - tracing the network traffic and quantifying the variation of the network traffic that flows into multiple outlets

Modeling - Seasonal Adjustment

Y_t	Original series
C_t	Trend component
S_t	Seasonal component
I_t	Irregular component
A_t	Adjusted series (without seasonal component)

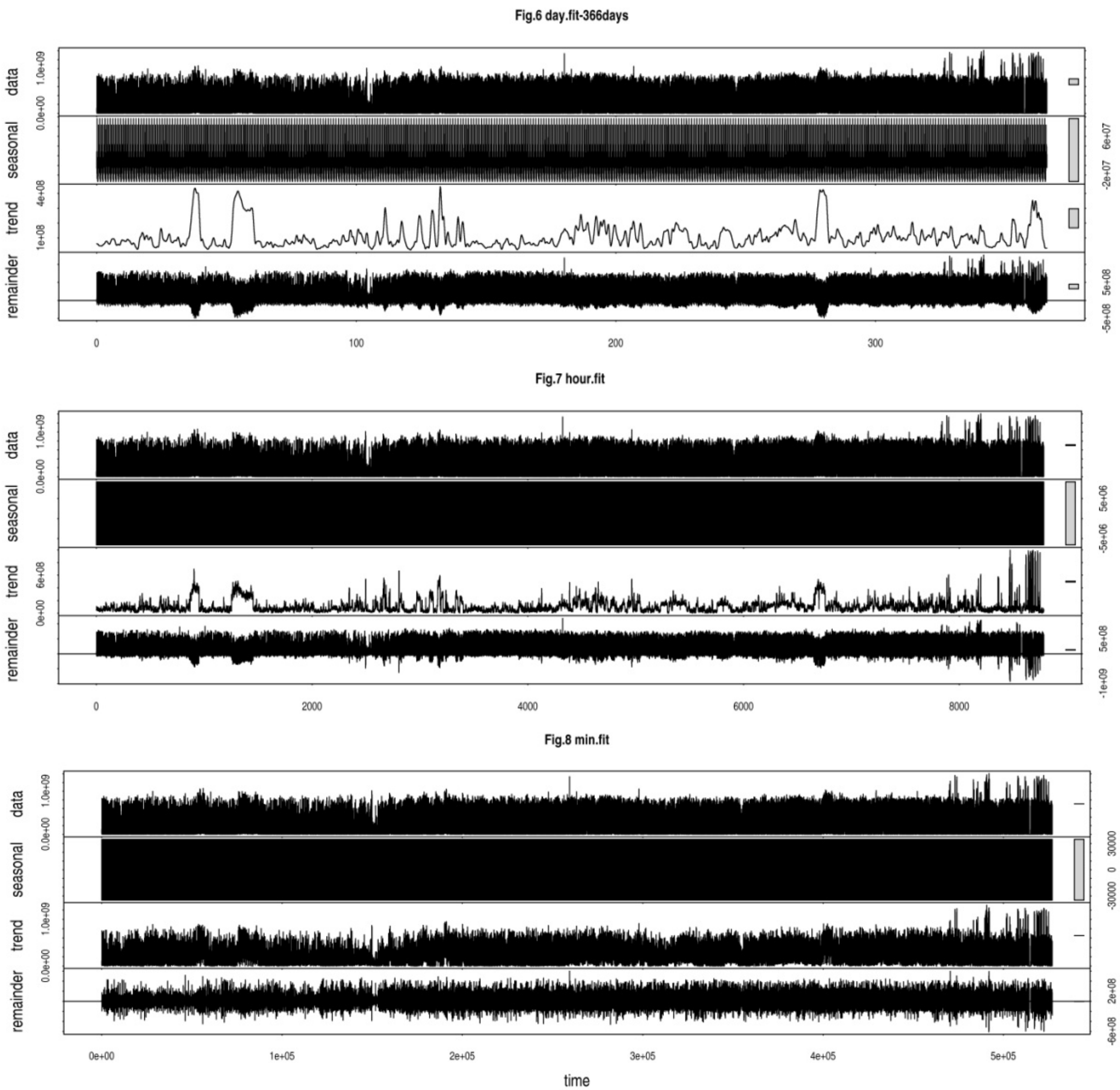
Multiplicative	$Y_t = S_t C_t I_t$	$A_t = C_t I_t$
Additive	$Y_t = S_t + C_t + I_t$	$A_t = C_t + I_t$
Log-Additive	$\mathrm{Ln}(Y_t) = C_t + S_t + I_t$	$A_t = \exp(C_t + I_t)$

Modeling - Seasonal Adjustment



- Seasonal adjustments showing original SNMP data and decomposed components with weekly periodicity at NERSC router

Statistical evaluation of periodicity



- **Seasonal adjustments showing original SNMP data and decomposed components with daily, hourly and minutely periodicity at NERSC router**

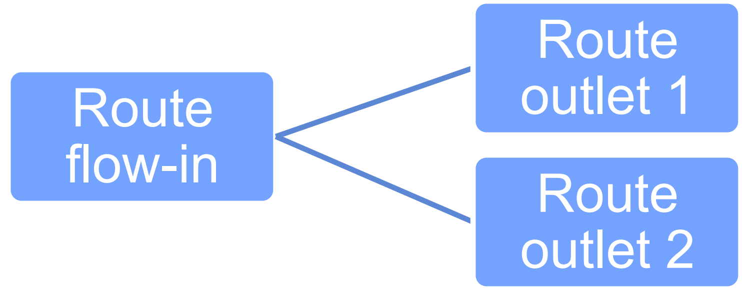
	Seasonal	Trend	Remainder
Weekly	43.2%	44.6%	78.3%
Daily	28.1%	69.1%	53.0%
Hourly	5.7%	88.5%	29.8%
Minute	0.1%	98.5%	6.8%

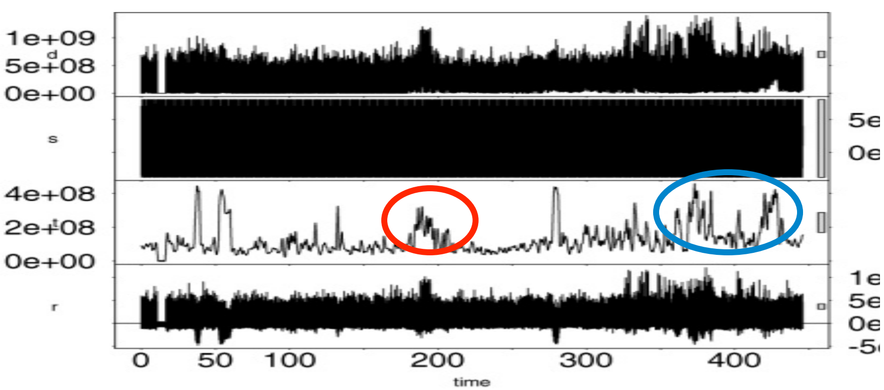
Evaluation of periodicity

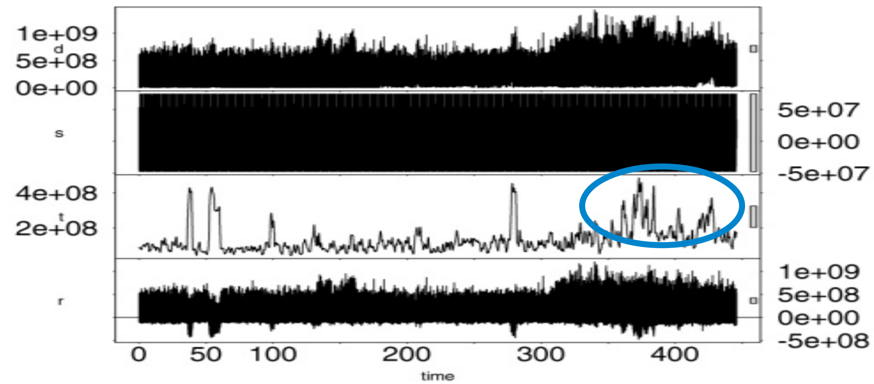
- Identified Seasonality
- Residual Seasonality
- Log transformation

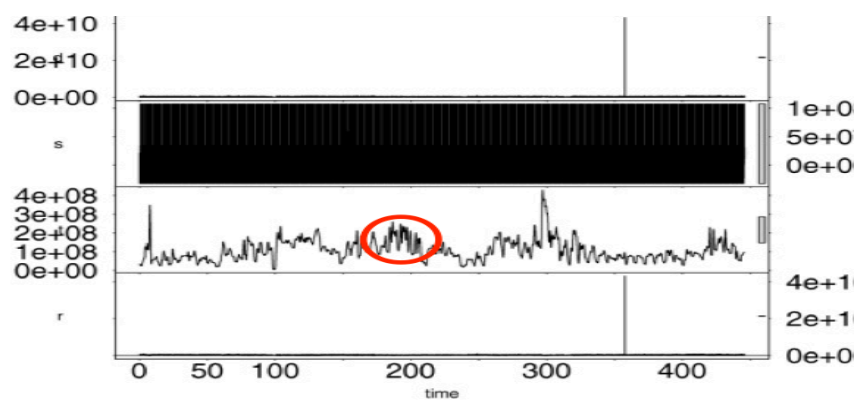
cycle	identified seasonality			residual seasonality			log transformation		
	Yes	no	Yes/total	yes	no	Yes/total	yes	no	Yes/total
Hour	59	119	0.331461	3	175	0.016854	128	50	0.719101
2hour	47	42	0.52809	3	86	0.033708	71	18	0.797753
8hour	6	16	0.272727	0	22	0	19	3	0.863636
10hour	2	15	0.117647	0	17	0	16	1	0.941176
12hour	6	8	0.428571	0	14	0	12	2	0.857143
14hour	1	11	0.083333	0	12	0	11	1	0.916667
20hour	1	7	0.125	0	8	0	7	1	0.875
22hour	0	8	0	0	8	0	7	1	0.875
day	7	0	1	0	7	0	7	0	1
2day	3	0	1	0	3	0	3	0	1
week	1	0	1	0	1	0	1	0	1

Tracing Data Flow

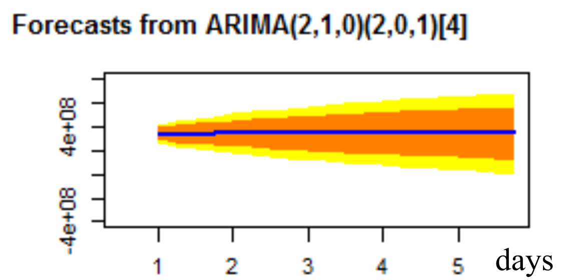


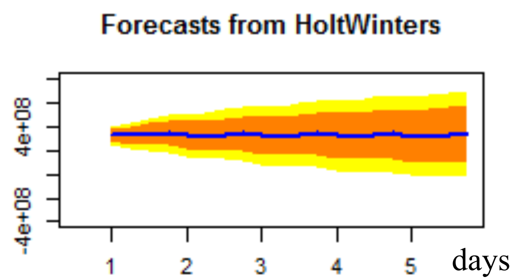


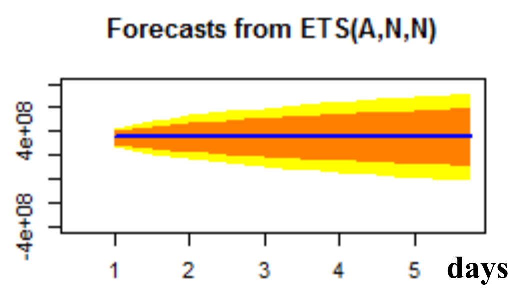


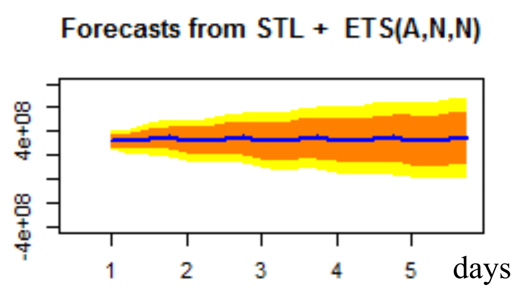


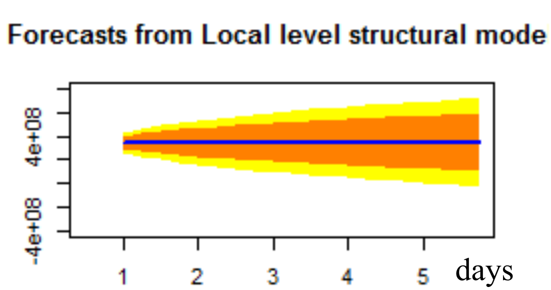
Performance of Prediction

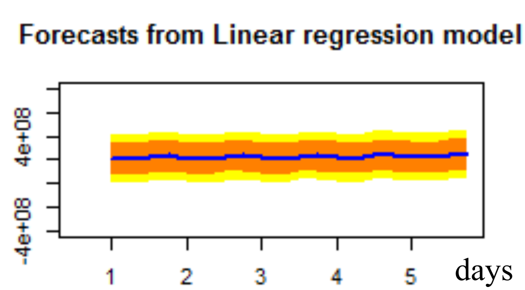










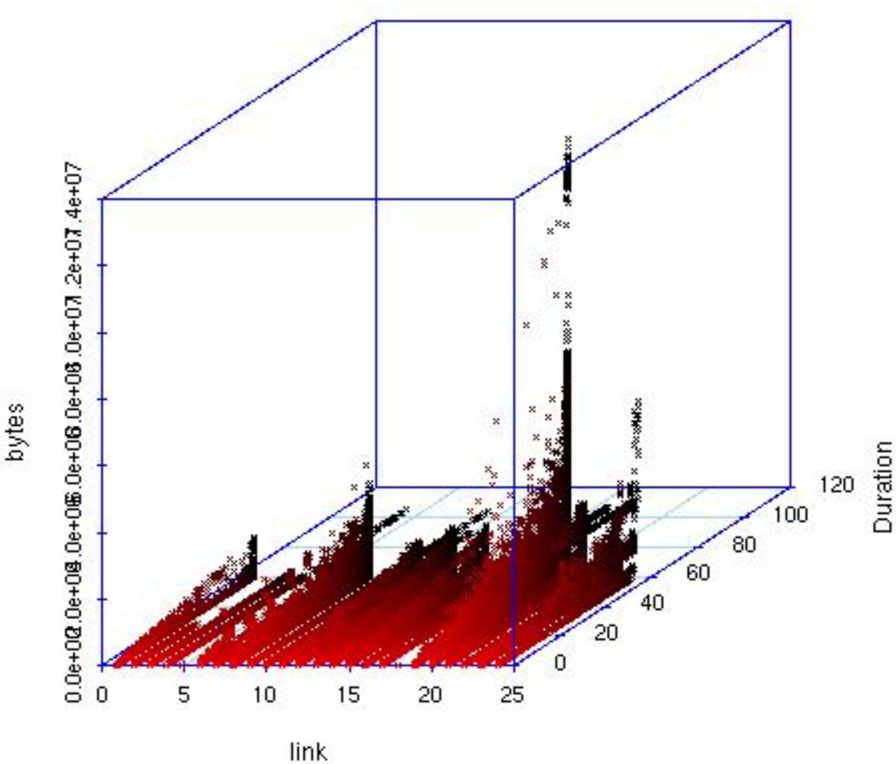


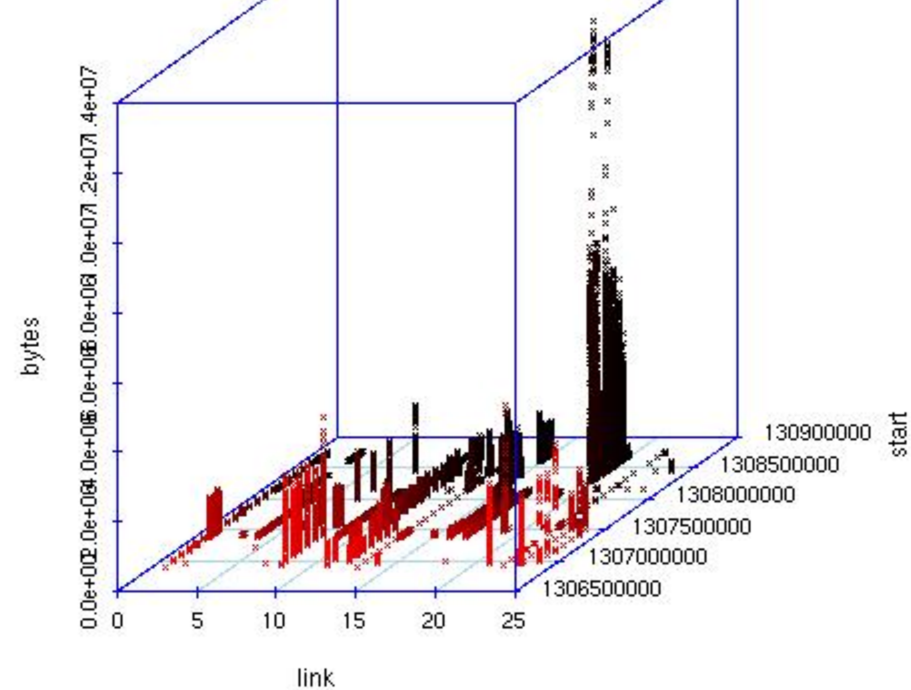
Generalized Linear Mixed Model

$$y = X\beta + Z\alpha + \epsilon$$

- For multivariate data analysis, the behavior of influence from each variable to the prediction is different.
  - Fixed effect: determined influence all the time
  - Random effect: random influence with a certain known distribution
- For multivariate data analysis, the randomness/ uncertainty comes from different sources
  - Error term: the universal variance caused by randomness
  - Random effect: the variance nested within the diversity in the network conditions

Netflow Data





3D Plots suggest random effects.

GLMM on Octets

- **How bytes is related to the timestamps and paths it take.**

$$y_{ij} = X\beta_j + Z\alpha_i + \epsilon_{ij} \quad i = 1, \ldots, m, j = 1, \ldots, l$$

where m is the number of distinct routes with direction

l is the number of unique timestamps

y_{ij} is the observational bytes

$X\beta_j$ is the mean of Octets for the jth observational time

$Z\alpha_i$ is a random effect associated with the ith route

ϵ_{ij} is an error term.

GLMM on Duration

- **How long to expect based on the size of the flow, start time of transfer and selected routes**

$$y_{ij} = X\beta_j + Z\alpha_i + \epsilon_{ij} \quad i = 1, \dots, m, j = 1, \dots, l$$

where m is the number of distinct routes with direction

l is the number of unique timestamps

y_{ij} is the observational duration

$X\beta_j$ is the mean of Duration for the certain size of flow at jth observation time

$Z\alpha_i$ is a random effect associated with the ith route

ϵ_{ij} is an error term.

GLMM on Octets

$$y_{ij} = X\beta_j + Z\alpha_i + \epsilon_{ij}$$

Octets ~ SrcP + DstIPaddress + Pkts

+ (1 | SrcIPaddress) + (Pkts |SrcIPaddress)

- **Fixed Effects: β_j Coefficient**

Intercept	54938.5706
SrcP (SourcePort)	-1.8182
DstIPaddress (Destination IP address)	4398.0490
Pkts(Packets)	3270.7419

- **Random Effects: α_i Variance**

SrcIPaddress (variance)	3.8210e+08
Pkts (variance)	1.0663e+07

GLMM on Duration

$$y_{ij} = X\beta_j + Z\alpha_i + \epsilon_{ij}$$

Duration ~ DstP + Pkts + start.r

+ (1 | SrcIPaddress) + (Pkts |SrcIPaddress)

- **Fixed Effects: β_j Coefficients**

Intercept	1.018e+02
DstP(Destination Port)	-2.252e-04
DstIPaddress (Destionation IP address)	4.152e-01
start.r	-7.597e-06

- **Random Effects: α_i Variance**

SrcIPaddress (variance)	123.39220
Pkts (variance)	0.17914

Conclusion

- Statistical approach to the prediction models for network traffic performance based on two types of data

- SNMP → Time series model with Seasonal Adjustment
  - Analyzing network traffic patterns
  - By decomposing into seasonal, trend and random components
  - To enable prediction, tracing and quantifying the network traffic
- Netflow → Generalized Linear Mixed Model
  - Analyzing variation with the network conditions
  - By considering fixed effects, random effects and error term
  - To improve accuracy of prediction by involving both universal varaince caused by randomness and variance by changes in the network traffic

More information

- Project web: http://sdm.lbl.gov/apm/
- Contact: kjhu@lbl.gov